VIII SEMINÁRIO DE PESQUISA EM ESTUDOS LINGUÍSTICOS
ISSN 2317-0549

COMPUTAÇÃO E LINGUÍSTICA: UM APLICATIVO WEB PARA BUSCAS AUTOMÁTICAS NO CORPUS DIGITAL DOVIC

Aline Silva Costa[97]
(UESB)

Cristiane Namiuti Temponi[98]
(UESB)

Jorge Viana Santos[99]
(UESB)

RESUMO

A descrição e compreensão dos aspectos gramaticais que figuraram no passado de uma determinada língua pressupõem desafios metodológicos aliados à vertente computacional das pesquisas em linguística. Nesta linha, o presente trabalho, desenvolvido no âmbito da Linguística de Corpus, objetiva apresentar o *status* do desenvolvimento de uma ferramenta Web que atenderá duplamente aos estudos linguísticos, disponibilizando o corpus digital DOViC na Internet e permitindo a realização de busca automática, baseada em categorias morfológicas e sintáticas. A disponibilidade da ferramenta contribuirá com a ampliação de pesquisas do português brasileiro e com a preservação do patrimônio linguístico e histórico de Vitória da Conquista.

[97] Professora do IFBa, mestranda do Programa de Pós-Graduação em Linguística. alinesilvacosta10@gmail.com
[98] Docente do Departamento de Estudos Linguísticos e Literários/Uesb. Orientadora.
[99] Docente do Departamento de Estudos Linguísticos e Literários/Uesb. Co-orientador.

ISSN 2317-0549

PALAVRAS-CHAVE: Aplicativo Web. Buscas-automáticas. Corpus DOViC.

INTRODUÇÃO

A questão central que se coloca para o trabalho com textos antigos como fundamentos para estudos linguísticos no meio eletrônico é, segundo Paixão de Sousa (2006), a busca por uma abordagem global do texto que se reflita numa integração entre diferentes planos de análise. Os estudos históricos realizados com base em textos antigos dependem da garantia da fidelidade às formas originais dos textos e, no caso dos *corpora* eletrônicos, esse pressuposto fundamental precisa ser integrado aos requerimentos impostos pela vertente computacional e linguística dos estudos – tais sejam: o arquivo virtual/digital, a confiabilidade e durabilidade do código, a necessidade de quantidade, agilidade e automação no trabalho de organização e seleção de dados. A Lingüística de *Corpus* ganha, assim, um papel central nos Estudos Diacrônicos. Nesta linha, propomo-nos a desenvolver ferramentas que possibilitem a recuperação instantânea de dados de língua em meio eletrônico, integradas a um aplicativo WEB.

ISSN 2317-0549

MATERIAL E MÉTODOS

Os textos do *corpus* digital DOViC (*Corpus* de Documentos Oitocentistas de Vitória da Conquista) são transcritos, editados e anotados nos mesmos moldes do *Corpus Tycho Brahe (CTB)*. O *CTB* é um *corpus* digital composto de textos em português de autores nascidos entre 1380 e 1845, desenvolvido na Universidade Estadual de Campinas, no âmbito do Projeto "Padrões Rítmicos, Fixação de Parâmetros e Mudança Linguística" (UNICAMP, 1998).

A transcrição e edição dos textos do *corpus* DOViC são feitos com o auxílio da ferramenta E-Dictor (KEPLER; PAIXÃO DE SOUSA; FARIA, 2010). O texto transcrito é salvo em um arquivo no formato texto simples. Edições como modernização, junção, segmentação e modernização de grafia são feitas por meio da interface gráfica da ferramenta, produzindo como resultado um arquivo anotado na linguagem XML. O software realiza anotação das informações morfológicas dos textos, também no formato XML e ambas as anotações são feitas num único arquivo. Esse esquema de anotação suportado pelo E-Dictor, utilizado tanto no *corpus* Tycho Brahe quanto no DOViC, foi concebido dentro do projeto "Memória dos Texto" (PAIXÃO DE SOUSA, 2006). Como esse

ISSN 2317-0549

processo é feito por meio da interface gráfica, o uso da linguagem XML é transparente para o usuário, ou seja, ele não lida diretamente com essa estrutura. A figura 1 mostra um trecho de um arquivo do *corpus* DOViC com as anotações de edição e morfológicas em XML, gerado pelo E-Dictor.

```
<text t="full" words="130" id="text_1">
  <sc id="sc_1">
    <p id="p_1">
      <s id="s_1">
        <w id="s_1#0">
          <o>eordeno</o>
          <e t="seg">e ordeno</e>
          <m v="CONJ"/>
          <m v="VB-P"/>
        </w>
        <w id="s_1#1">
          <o>atodos</o>
          <e t="seg">a todos</e>
          <m v="P"/>
          <m v="Q-P"/>
        </w>
        <w id="s_1#2">
          <o>osOfficiaes</o>
          <e t="mod">os oficiais</e>
          <e t="seg">os Officiaes</e>
          <m v="D-P"/>
          <m v="N-P"/>
        </w>
        <w id="s_1#3">
          <o>de</o>
          <m v="P"/>
        </w>
        <w id="s_1#4">
          <o>Justiça</o>
          <m v="NPR"/>
        </w>
        <w id="s_1#5">
          <o>desta</o>
          <m v="P+D-F"/>
        </w>
        <w id="s_1#6">
          <o>sobre</o>
```


Figura 1: Arquivo XML gerado pelo E-Dictor para um documento do *corpus* DOViC

O aplicativo Web, em fase de programação, segue o processo de desenvolvimento de software iterativo e incremental. Na implementação fez-se uso das seguintes tecnologias: *Framework JSF* (*Java Server Faces*); *Unified Modeling Language* – UML e Sistema de

ISSN 2317-0549

Banco de Dados *PostgreSQL*. Para as buscas automáticas, utiliza-se as tecnologias para a linguagem XML: XSLT (*eXtensible StyleSheet Language Transformation*) e XQuery.

RESULTADOS E DISCUSSÃO

O aplicativo Web para disponibilização do corpus DOViC na Internet está em fase de desenvolvimento e diversas funcionalidades já foram implementadas. Por se tratar de um *corpus* de documentos históricos manuscritos, as fotografias dos documentos originais do DOViC também são disponibilizadas no aplicativo. No *status* atual, já é possível armazenar os manuscritos do corpus no banco de dados e visualizar suas principais características, como o texto transcrito, as imagens dos originais e as informações gerais (título, local de depósito, data, gênero, informações sobre edição, etc). A figura 2 mostra a tela do aplicativo exibindo dados de um livro de escrituras que compõem o *corpus* DOViC. A figura 3 mostra uma tela com as informações de uma carta de alforria do *corpus*. Todas estas informações e imagens já estão armazenadas no banco de dados, do qual o aplicativo recupera as informações.

VIII SEMINÁRIO DE PESQUISA EM ESTUDOS LINGUÍSTICOS
ISSN 2317-0549

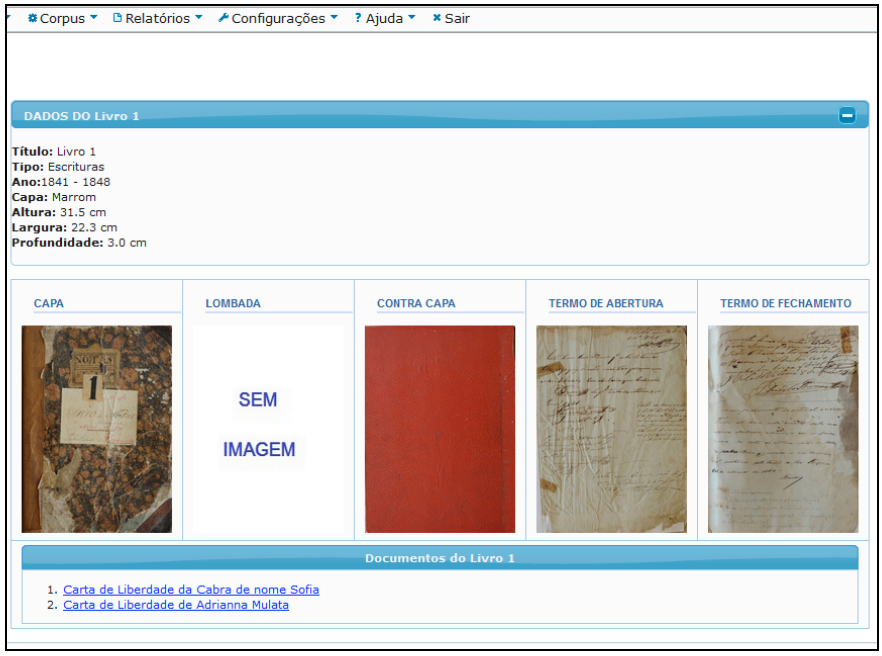

Figura 2: Tela do aplicativo Web exibindo informações do livro 1 do *corpus* DOViC

ISSN 2317-0549

Figura 3: Tela do aplicativo Web exibindo carta de alforria do *corpus* DOViC

A ferramenta de busca que deve compor o aplicativo web que possibilitará pesquisas linguísticas através de uma interface gráfica amigável está em fase de projeto. As buscas morfológicas serão feitas no arquivo XML gerado pelo E-Dictor. A versão atual deste programa (versão 1.0 beta 10) não realiza anotação da estrutura sintática. Tal informação é gerada separadamente utilizando um *parser* que recebe como entrada um arquivo anotado no formato POS (*Part of Speech*), e gera como saída um arquivo texto no formato

ISSN 2317-0549

Penn TreeBank[100]. O treinamento do *parser* foi feito para o português clássico na Universidade da Pensilvânia. Para obtenção da representação sintática nos textos do *corpus* DOViC, os textos deverão ainda passar pelo mesmo processo de etiquetagem. As buscas sintáticas serão feitas nesses arquivos, mas com uso de uma interface gráfica, tornando todo o esquema de anotação transparente para o usuário.

CONCLUSÕES

O aplicativo web proposto no trabalho encontra-se na fase de desenvolvimento. A ferramenta de busca está em fase de projeto e deverá compor o aplicativo, permitindo buscas automáticas para auxílio de pesquisas linguísticas através de uma interface gráfica amigável. A disponibilização do *corpus* digital *DOViC* na Internet irá contribuir com a ampliação de pesquisas sobre o português brasileiro e também com a preservação do patrimônio linguístico e histórico da cidade de Vitória da Conquista.

[100] O *Penn TreeBank Format* é um esquema de anotação sintática de *corpora* desenvolvido pela Universidade da Pensilvânia. O esquema utiliza uma representação arbórea delimitada por parênteses etiquetados.

ISSN 2317-0549

REFERÊNCIAS

PAIXÃO DE SOUSA, M.C. Memórias do Texto. **Revista Texto Digital**, n.2., 2006. Disponível em: <http://www.textodigital.ufsc.br/num02/paixao.htm>. Acesso em: 01 out 2013.

SANTOS, J.V. (Coord.) **Memória Conquistense: recuperação de documentos oitocentistas na implementação de um *corpus* digital**. Projeto de Pesquisa. UESB, Vitória da Conquista, 2009.

UNICAMP. ***Corpus* Histórico Anotado do Português Tycho Brahe**. 1998. Disponível em: <www.tycho.iel.unicamp.br/~tycho/corpus>. Acesso em: 01 ago 2013.